Linear Regression

Huy Quang Lai
Department of Computer Science and Engineering
Texas A&M University
College Station, Texas 77843
lai.huy@tamu.edu

1 Introduction

Linear regression is a fundamental statistical technique used to establish a relationship between two or more variables. Linear regression is a simple yet powerful method used to predict a dependent variable's value based on the values of one or more independent variables. As its name implies, linear regression models can analyze the linear relationship between a dependent variable and one or more independent variables. It is widely used in various fields, including finance, economics, social sciences, and engineering, to make predictions and inform decision-making processes. Linear regression analysis provides valuable insights into the underlying data and can help identify trends, patterns, and outliers.

2 Defining a Linear Regression

Given a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ where $\boldsymbol{x}_i$ is a vector of length D and y_i is a scalar.

Using $\mathcal{D}$ as the data set for a linear regression model, the linear regression is defined as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_D x_{i,D} + \varepsilon_i, i = 1, \cdots, N$$

Or in vector form,

$$y = \beta \boldsymbol{X}_i^T + \varepsilon, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}, X = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,D} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_D \end{bmatrix}$$

This equation represents the true relationship between the dependent variables and the independent variable.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

3 Simple Linear Regression

A simple linear regression is a type of linear regression where there is only one independent variable. It is defined to be

$$y = \beta_0 + \beta_1 x + \varepsilon, \varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, 2, \cdots, N$$

The parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be derived by minimizing the residual sum of squares.

3.1 Optimal Coefficients

Proof.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sigma(xy)}{\sigma(x)^2}$$

The Residual Sum of Squares is defined as

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

Taking the partial derivative of RSS with respect to β_0 and β_1 results in

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{N} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2)$$

To minimize these two variables, set both partial derivatives to zero.

$$-2 \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$-2 \sum_{i=1}^{N} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$$

This yeilds the following system of equations

$$\hat{\beta}_1 \sum_{i=1}^{N} x_i + \hat{\beta}_0 N = \sum_{i=1}^{N} y_i$$
$$\hat{\beta}_1 \sum_{i=1}^{N} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i$$

From the first equation, the estimate for the intercept can be derived:

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{N} \sum_{i=1}^{N} y_i - \hat{\beta}_1 \frac{1}{N} \sum_{i=1}^{N} x_i \\ &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

From the second equation, the estimate for the slope can be derived:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{N} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} x_i y_i \\
\hat{\beta}_1 \sum_{i=1}^{N} x_i^2 + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} x_i y_i \\
\hat{\beta}_1 \left(\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i \right) &= \sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i}
\end{aligned}$$

Note that the numerator can be rewritten as

$$\begin{aligned}
\sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y} \\
&= \sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i - \bar{x} \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \bar{x}\bar{y} \\
&= \sum_{i=1}^{N} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y_i})
\end{aligned}$$

and that the denominator can be rewritten as

$$\begin{aligned}
\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{N} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{N} x_i^2 - 2\bar{x} \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \bar{x}^2 \\
&= \sum_{i=1}^{N} (x_i^2 - 2\bar{x} x_i + \bar{x}^2) \\
&= \sum_{i=1}^{N} (x_i - \bar{x})^2
\end{aligned}$$

Substituting these results into the original expression for $\hat{\beta}_1$ results in

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y_i})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \\
&= \frac{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y_i})}{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \\
&= \frac{\sigma(xy)}{\sigma(x)^2}
\end{aligned}$$

Therefore, the optimal parameters for a simple linear regression are,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sigma(xy)}{\sigma(x)^2}$$

□

4 Least Absolutes

The residual sum of absolutes, or RSA, is defined as

$$RSA = \sum_{i=1}^{N} \left| y_1 \hat{\beta}_0 - \hat{\beta}_1 X_{1,N} \right|$$

4.1 Advantages over Least Squares

When an error is over 1, squaring it will cause the error to increase. Additionally, when outliers exist, least squares becomes more susceptible to them and potentially overfits the model.

As an example, of how least squares can overfit to outliers is displayed in the figure[1] below.

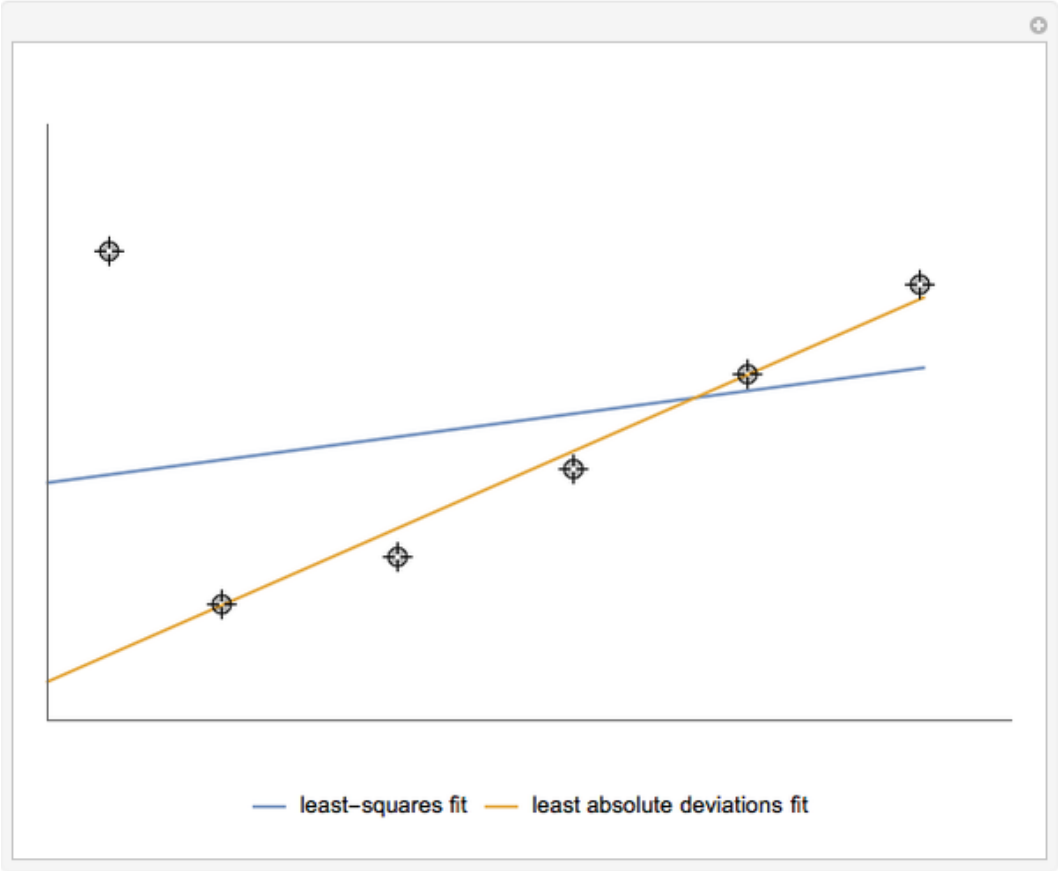

Figure 1: Least Squares vs Least Absolutes

[1] Image source here

5 Accuracy of Coefficient Estimates

5.1 The Question of Coefficients

Assume that the true relationship of a simple linear regression is

$$y = \beta_0 + \beta_1 x + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

However, when creating a linear regression, only a sample of the total population is used.

For example, the following image[2] shows two possible linear regression models with a sample size of 100.

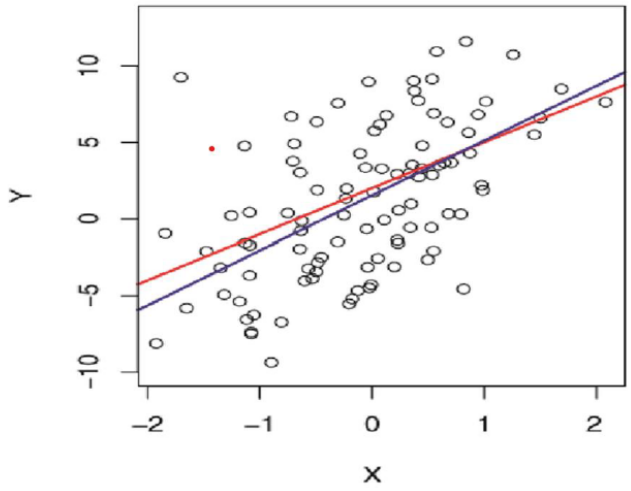

Figure 2: Two possible Simple Linear Regression Models

Continuing the process for many simple linear regression models can result in the following graph[3].

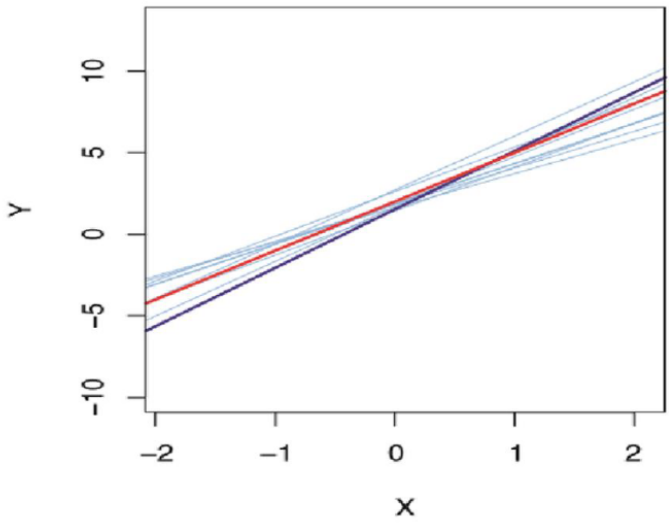

Figure 3: Many possible Simple Linear Regression Models

[2]Image source: lecture slides
[3]Image source: lecture slides

With the numerous possible simple linear regression models, a question arrises: which definition of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\mu}$ is best?

5.2 Finding the Optimal Parameters

The Standard Error can assist in determining how much these parameters deviate from their true values.

$$\begin{aligned} SE\left(\hat{\mu}\right)^2 &= \frac{\sigma^2}{N} \\ SE\left(\hat{\beta}_0\right)^2 &= \sigma^2\left(\frac{1}{N} + \frac{x^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right) \\ SE\left(\hat{\beta}_1\right)^2 &= \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ \sigma^2 &= Var(\varepsilon) \end{aligned}$$

When x_i is smaller values that are spread out, more leverage is required to estimate the slope. This will reduce $SE\left(\hat{\beta}_1\right)$. Additionally, when $\bar{x} = 0$, $SE\left(\hat{\beta}_0\right) = SE\left(\bar{\mu}\right)$

σ itself, however is not knows. But a good estimate of its value is the residual standard error

$$RSE = \sqrt{\frac{RSS}{N-2}}$$

6 Hypothesis Testing

6.1 The Hypothesis

The use of Standard Error allows for hypothesis testing. A very common hypothesis is the Null Hypothesis and its alternative hypothesis.

H_0 : There is no relationship between x and y
H_a : There exists a relationship between x and y

Mathematically this is equivalent to

$$\begin{aligned} H_0 &: \beta_1 = 0, \therefore y = \beta_0 + \varepsilon \\ H_a &: \beta_1 \neq 0, \therefore \hat{\beta}_1 >> 0 \end{aligned}$$

6.2 T-statistic

$$t_\beta = \frac{\hat{\beta}_1 - \beta}{SE\left(\hat{\beta}_1\right)}$$

$$t_\beta = \frac{\hat{\beta}}{SE\left(\hat{\beta}_1\right)} \text{ for } H_0$$

If no relationship between x and y exists, the data is expected to be a t-distribution with P-2 degrees of freedom. Compute the probability of observing any number equal to $|t|$ or larger in absolute value, assuming $\beta_1 = 0$. This probability is called the p-value. For a small p-value, is unlikely to observe a substantial association between predictor and response due to chance. Therefore, a small p-value means there is an association between x and y so we can reject the null hypothesis. The cutoff is usually 5% or 1%

Table 1: Data for the Example

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

6.3 Example

With the sample size of 30, the t-statistic for the null hypothesis of both the Intercept and TV is about 2 and 2.75 respectively. With this result, $\beta_0 \neq 0$ and $\beta_1 \neq 1$

7 Accuracy of the Simple Linear Regression

Once the null hypothesis for β_0 and β_1 is rejected, a natural follow-up question is how well the model fits the data.

7.1 Residual Standard Error

One measure is the residual standard error

$$RSE = \sqrt{\frac{RSS}{N-2}}$$

However, it is not always clear what a good value of RSE is.
Another possible measure is the Coefficient of Determination.

7.2 The Coefficient of Determination

The Coefficient of Determination is defined as

$$R^2 = \frac{SEE}{TSS}$$

Where the residual sum of squares and the total sum of squares, or SSE and TSS respectively, is defined as

$$RSS = \sum_{i=1}^{N} (\hat{y} - y_i)^2$$
$$TSS = \sum_{i=1}^{N} (\bar{y} - y_i)^2$$

Proportion of variance explained, always between 0 and 1, independent of scale of y. The total sum of squares measures the total variance in response y.
The residual sum of squares accounts for the remaining error left unexplained after the regression.

When R^2 is close to 1, a large proportion of variation is explained by the regression. In contrast, when R^2 is close to 0, the regression does not explain the variation. Perhaps the model is wrong or σ^2 is too large. R^2 is a measure of the relationship between x and y.

8 Correlation Coefficient in Simple Linear Regression

The correlation of two random variables X and Y, also called Pearson product-moment correlation coefficient, is defined as

$$r = \frac{\sigma(xy)}{\sigma(x)\sigma(y)}$$

An alternative definitions for the correlation coefficient is

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y - \bar{y})^2}}$$

8.1 Relationship with the Slope Estimate

In simple linear regression the correlation coefficient is related to the slope of the simple linear regression.

$$r = \frac{\sigma(x)}{\sigma(y)}\hat{\beta}_1$$

Proof. Using the definition of correlation,

$$r = \frac{\sigma(xy)}{\sigma(x)\sigma(y)}$$
$$\sigma(xy) = \sigma(x)\sigma(y)r$$

Substituting this derivation into the definition of $\hat{\beta}_1$ derived in Proof 3.1 results in

$$\hat{\beta}_1 = \frac{\sigma(x)\sigma(y)r}{\sigma(x)^2}$$
$$= \frac{\sigma(y)}{\sigma(x)}r$$
$$\Leftrightarrow r = \frac{\sigma(x)}{\sigma(y)}\hat{\beta}_1$$

□

8.2 Relationship with the Coefficient of Determination

In simple linear regression the correlation coefficient is related to the coefficient of determination, or R^2.

$$r^2 = R^2$$

Proof. Recall from Proof 3.1 that the optimal definitions of the simple linear regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$
$$\hat{\beta}_1 = \frac{\sigma(xy)}{\sigma(x)^2}$$

Additionally, recall that coefficient of determination, R^2, is defined as

$$R^2 = \frac{SSE}{TSS}$$

Using the definition of the coefficient of determination,

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y} - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}$$
$$= \frac{\sum_{i=1}^{N}(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Substituting the definition of $\hat{\beta}_0$ results in

$$\begin{aligned}
R^2 &= \frac{\sum_{i=1}^{N}(\bar{y} - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{N}(\hat{\beta}_1(\bar{x} - x_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\sum_{i=1}^{N}(\bar{x} - x_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\sigma(x)^2}{\sigma(y)^2} \\
&= \left(\hat{\beta}_1 \frac{\sigma(x)}{\sigma(y)}\right)^2
\end{aligned}$$

Recall from Proof 8.1 that

$$r = \frac{\sigma(x)}{\sigma(y)}\hat{\beta}_1$$

Using this relationship,

$$R^2 = r^2$$

□

References

[1] Watt, Jeremy, Borhani, Reza & Katsaggelos, Aggelos Konstantinos (2016) Machine Learning Refined.

[2] Konasani, Venkata Reddy & Shailendra Kadre (2021) Machine Learning and Deep Learning Using Python and TensorFlow.

[3] Simple Linear Regression

[4] Coefficient of Determination

[5] Correlation